

GIAB data submission and distribution: an overview

Chunlin Xiao, NCBI

BioProject: PRJNA200694 for Genome-in-a-Bottle

NCBI

Resources

How To

BioProject

BioProject

200694[uid]

Create alert

Advanced

Display Settings

Send to

Genome in a Bottle (human)

Accession: PRJNA200694 ID: 200694

A public-private-academic consortium hosted by NIST to develop reference materials and standards for clinical sequencing

The Genome in a Bottle Consortium (www.genomeinabottle.org) is a collaboration between NIST, FDA, NCBI, other government agencies, academic sequencing groups, sequencing technology developers, and clinical laboratories. A principal motivation for this consortium is to develop widely accepted reference materials and accompanying performance metrics to provide a strong scientific foundation for the development of regulations and professional standards for clinical sequencing. NIST is developing large batches of human genome DNA from several cell lines for NIST Reference Materials (RMs), which will be characterized by the Consortium for homogeneity, stability, and sequence with as much sequencing technologies and library preparation methods as possible. Information from these datasets will be integrated to form a high-confidence set of genotype calls, which can be used by clinical and research laboratories to understand performance of their sequencing and bioinformatics methods. NCBI is serving as the DCC and repository for the raw sequencing reads, mapped reads, genotypes, and other details for each sample on a dedicated FTP site (<ftp://ftp-trace.ncbi.nih.gov/giab/ftp/>). The pilot sample is NA12878 (HG001), and NIST received over 8,000 aliquots in April 2013, which will initially be distributed to partners in the Consortium to assist in characterization, and later will be distributed by NIST as Reference Material 8398, likely in March or April 2015. Samples from an Ashkenazim trio (son HG002-NA24385-huAA53E0, father HG003-NA24149-hu6E4515, and mother HG004-NA24143-hu8E87A9), and a Han Chinese trio (son HG005-NA24631-hu91BD69, father NA24694-huCA017E, and mother NA24695-hu38168C) from Personal Genome Project (PGP) are also candidate NIST reference materials and are currently being characterized. The Ashkenazim trio will be available both as NIST RMs 8391 (son only) and 8392 (entire trio). Only the son of the Asian trio will be a NIST RM (8393). DNA and cell lines for all samples are also available from Coriell, but the NIST RMs are from a single homogenized batch of DNA, so there may be small differences between the samples at Coriell and the NIST RMs. Details about the NIST Reference Materials, data, and future plans are at <https://sites.stanford.edu/abms/content/giab-reference-materials-and-data>. When the NIST RMs are available, they can be purchased from NIST at <http://www.nist.gov/srm/>, where a Report of Investigation describing the DNA will also be available. [Less...](#)

See [Genome](#) Information for [Homo sapiens](#)

Navigate Across
29876 additional projects are related by organism.

- NA12878 (HG001)
- Ashkenazim trio (HG002, HG003, HG004)
- Chinese Trio (HG005, HG006, HG007)

Accession	PRJNA200694
Data Type	Genome sequencing and assembly
Scope	Multiisolate
Organism	Homo sapiens [Taxonomy ID: 9606] Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo; Homo sapiens
Submission	Registration date: 28-Jan-2016 Genome in a Bottle <ul style="list-style-type: none">- National Institute of Standards and Technology- National Center for Biotechnology Information
Related	<ul style="list-style-type: none">• GiaB home page
Resources	<ul style="list-style-type: none">• GiaB ftp• GiaB:GitHub

Project data summary for Genome-in-a-Bottle (under PRJNA200694)

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (total)	5
WGS master	3
SRA Experiments	764
OTHER DATASETS	
BioSample	17
Assembly	3

Assembly details:

Assembly level					Number of Assemblies	
Contig					3	
Total					3	
Assembly	Level	WGS	Chrs	BioSample	Isolate	Taxonomy
GCA_001542345.1		LRIL000000000		SAMN03283347	NA24385	Homo sapiens
GCA_001549595.1		LRUM000000000	1	SAMN03283346	NA24143	Homo sapiens
GCA_001549605.1		LRUL000000000	1	SAMN03283345	NA24149	Homo sapiens

SRA Data Details

Parameter	Value
Data volume, Gbases	9,361
Data volume, Tbytes	8.62

Links from BioProject

Items: 17

☐

[whole exome sequencing of HG005](#)

1. Identifiers: BioSample: SAMN04299543; Sample name: HG005; SRA: SRS1181289
Organism: Homo sapiens
Isolate: NA24631
Package: Human; version 1.0
Accession: SAMN04299543 ID: 4299543
[BioProject](#) [SRA](#)

☐

[whole exome sequencing of HG004](#)

2. Identifiers: BioSample: SAMN04299542; Sample name: HG004; SRA: SRS1181290
Organism: Homo sapiens
Isolate: NA24143
Package: Human; version 1.0
Accession: SAMN04299542 ID: 4299542
[BioProject](#) [SRA](#)

☐

[Whole exome sequencing of HG003](#)

3. Identifiers: BioSample: SAMN04299541; Sample name: HG003; SRA: SRS1181288
Organism: Homo sapiens
Isolate: NA24149
Package: Human; version 1.0
Accession: SAMN04299541 ID: 4299541
[BioProject](#) [SRA](#)

☐

[Whole exome sequencing of HG002](#)

4. Identifiers: BioSample: SAMN04299540; Sample name: HG002; SRA: SRS1181267
Organism: Homo sapiens
Isolate: NA24385
Package: Human; version 1.0
Accession: SAMN04299540 ID: 4299540
[BioProject](#) [SRA](#)

Links from BioProject

Items: 3

☐

[GIAB Ashkenazim Son HG002/NA24385/huAA53E0 PacBio Assembly with PBcR](#)

1. Organism: Homo sapiens (human)
Sex: male
Submitter: Genome in a Bottle
Date: 2016/01/28
Assembly level: Contig
Genome representation: full
GenBank assembly accession: GCA_001542345.1 (latest)
RefSeq assembly accession: n/a
IDs: 632441 [UID] 2851158 [GenBank]

☐

[GIAB Ashkenazim Mother HG004/NA24143/huE87A9 PacBio Assembly with PBcR](#)

2. Organism: Homo sapiens (human)
Sex: female
Submitter: Genome in a Bottle
Date: 2016/02/16
Assembly level: Contig
Genome representation: full
GenBank assembly accession: GCA_001549595.1 (latest)
RefSeq assembly accession: n/a
IDs: 642201 [UID] 2879138 [GenBank]

☐

[GIAB Ashkenazim Father HG003/NA24149/hu6E4515 PacBio Assembly with PBcR](#)













3. Organism: Homo sapiens (human)
Sex: male
Submitter: Genome in a Bottle
Date: 2016/02/16
Assembly level: Contig
Genome representation: full
GenBank assembly accession: GCA_001549605.1 (latest)
RefSeq assembly accession: n/a
IDs: 642191 [UID] 2879118 [GenBank]

Official GIAB ftp site: structure/organization

Index of <ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/>

 Up to higher level directory

Name

 [CHANGELOG](#)
 [README.ftp_structure](#)
 [README.s3_structure](#)
 [changelog_details](#)
 [current.tree](#)
 [data](#)
 [data_indexes](#)
 [giab_s3_urls](#)
 [release](#)
 [technical](#)
 [tools](#)
 [use_cases](#)

Size

1 KB
4 KB
3 KB

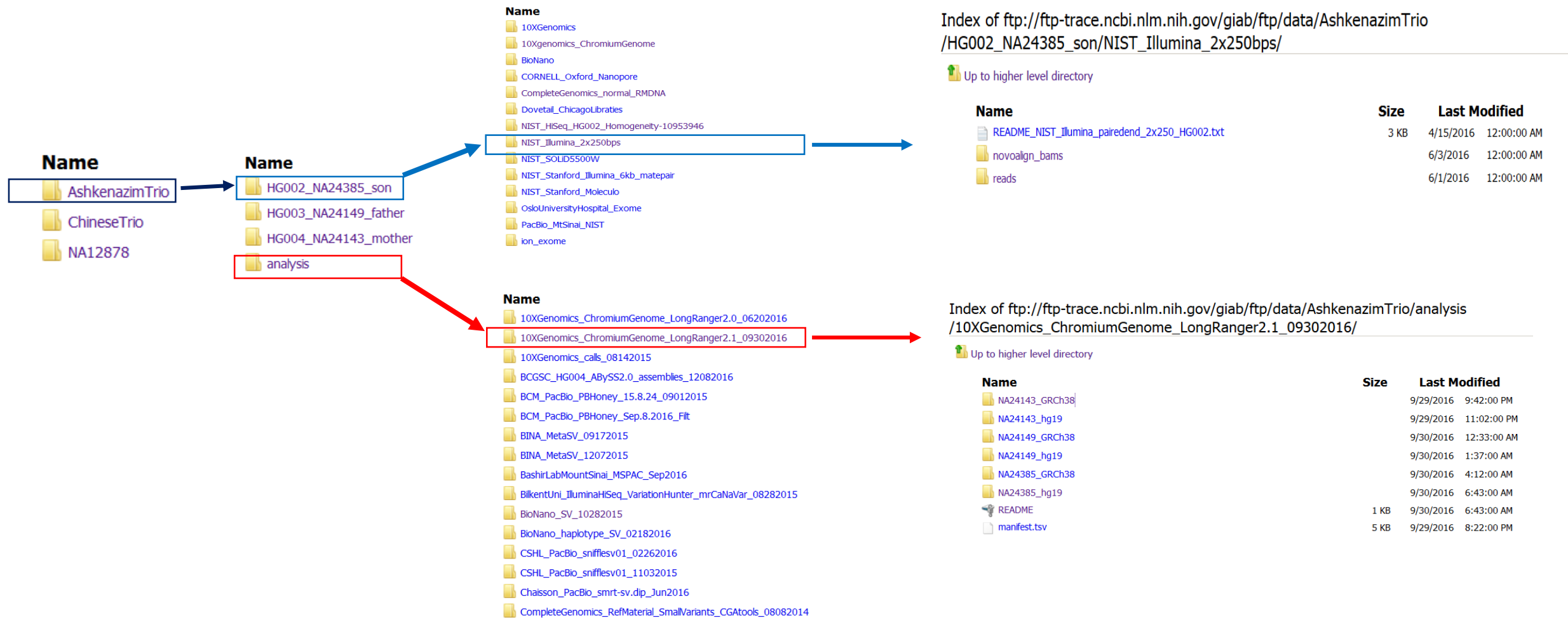
17124 KB

16785 KB

Last Modified

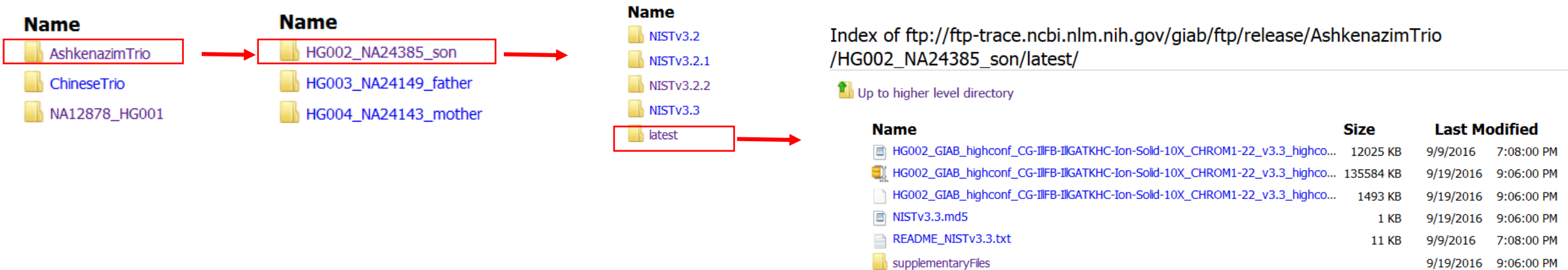
9/11/2015 12:00:00 AM
10/1/2015 12:00:00 AM
7/1/2016 12:00:00 AM
9/11/2015 12:00:00 AM
12/8/2016 5:43:00 PM
10/5/2014 12:00:00 AM
3/4/2016 12:00:00 AM
12/8/2016 5:43:00 PM
9/9/2016 7:48:00 PM
10/5/2014 12:00:00 AM
1/15/2014 12:00:00 AM
10/5/2016 8:14:00 PM

Under ftp/data: reads, bams, analysis results (variant callsets, assemblies etc)




*under "analysis", directory name convention follows the rule of: Institute_name + Platform + Analysis_method + Date

Under ftp/release: official callsets with versions by samples



GitHub site - <https://github.com/genome-in-a-bottle>



The screenshot shows the GitHub repository page for 'genome-in-a-bottle'. The repository is described as 'A public-private-academic consortium hosted by NIST to develop reference materials and standards for clinical sequencing'. It features a search bar, filters for 'Type' and 'Language', and a list of repositories. Two red arrows point to the 'giab_data_indexes' and 'giab_publications' repositories. The 'giab_data_indexes' repository is described as 'This repository contains data indexes from NIST's Genome in a Bottle project.' and has 27 stars and 7 forks. The 'giab_publications' repository is described as 'This repository contains a list of publications pertinent to the Genome in a Bottle project' and has 3 stars and 1 fork. The 'People' section on the right indicates that the organization has no public members.

genome-in-a-bottle
A public-private-academic consortium hosted by NIST to develop reference materials and standards for clinical sequencing

Repositories **People** 0

Search repositories... Type: All Language: All

giab_data_indexes
This repository contains data indexes from NIST's Genome in a Bottle project.
★ 27 🍴 7 Updated on Oct 14, 2016

giab_announcement_blog
This repository contains recent announcement or blog post from the Genome in a Bottle Consortium
GCC Machine Description ★ 3 🍴 1 Updated on Sep 12, 2016

giab_latest_release
This repository contains information about latest release from Genome in a Bottle project
★ 9 🍴 1 Updated on Sep 12, 2016

giab_publications
This repository contains a list of publications pertinent to the Genome in a Bottle project
★ 3 🍴 1 Updated on Jun 8, 2016

giab_data_analysis
This repository contains information about ongoing analysis performed by GIAB
★ 4 🍴 1 Updated on Mar 31, 2016

Top languages
GCC Machine Description

People 0 >
This organization has no public members. You must be a member to see who's a part of this organization.

giab_data_indexes

This repository contains data indexes from NIST's Genome in a Bottle project. The indexes for sequences and alignments are also under: ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data_indexes/.

AshkenazimTrio

Sequencing Platform	Sequence Index or Alignment Index
Illumina WGS 2x150bp 300X per individual	sequence.index.AJtrio_Illumina300X_wgs_07292015
	alignment.index.AJtrio_Illumina300X_wgs_novoalign_GRCh37_GRCh38_NHGRI_07282015
Illumina 6KB Matepair	sequence.index.AJtrio_Illumina_6kb_matepair_wgs_08032015
	alignment.index.AJtrio_Illumina_6kb_matepair_wgs_bwamem_GRCh37_07302015
	sequence.index.AJtrio_Illumina_2x250bps_06012016
	alignment.index.AJtrio_Illumina_2x250bps_isaac-align_hg19_06012016
	alignment.index.AJtrio_Illumina_2x250bps_novoalign_GRCh37_GRCh38_NHGRI_06062016
Moleculo	sequence.index.AJtrio_NIST_Stanford_Moleculo_125bps_08042015
PacBio 70x/30x/30x	sequence.index.AJtrio_PacBio_MtSinai_NIST_hdf5_08072015 , alignment.index.AJtrio_PacBio_MSSM_blasr_GRCh37_11192015
	alignment.index.AJtrio_PacBio_CSHL_bwamem_GRCh37_11192015
Oxford Nanopore	sequence.index.AJtrio_HG002_Cornell_Oxford_Nanopore_fasta_fastq_10132015
SOLID 60x for son	sequence.index.AJtrio_HG002_NIST_SOLiD5500W_xsq_09042015
	alignment.index.AJtrio_HG002_SOLiD5500W_NIST_LifeScope_GRCh37_12212015
Illumina Whole Exome by Oslo Uni. Hospital	alignment.index.AJtrio_OsloUniversityHospital_IlluminaExome_bwamem_GRCh37_11252015
Ion Proton 1000x Exome	alignment.index.AJtrio_IonTorrent_exome_TMAP_GRCh37_07292015
10X Genomics	alignment.index.AJtrio_10XGenomics_bwamem_GRCh37_08142015
10X Genomics ChromiumGenome	alignment.index.AJtrio_10Xgenomics_ChromiumGenome_GRCh37_GRCh38_06202016
CompleteGenomics	alignment.index.AJtrio_CompleteGenomics_normal_RMDNA_EvidenceBams_GRCh37_09282015
CompleteGenomics LFR	CG LFR raw or alignment data not available, but analysis results available under: ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/CompleteGenomics_newLFR_CGAtools_06122015/
BioNano	sequence.index.AJtrio_BioNano_bnx_10012015
	alignment.index.AJtrio_BioNano_xmap_cmap_GRCh37_10012015

89 lines (88 sloc) | 29.6 KB

RawBlameHistory

```
1 FASTQ FASTQ_MD5 PAIRED_FASTQ PAIRED_FASTQ_MD5 NIST_SAMPLE_NAME
2 ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG004_NA24143_mother/NIST_Illumina_2x250bps/reads/D3_S1_L001_R1_001.fastq.gz
3 ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG004_NA24143_mother/NIST_Illumina_2x250bps/reads/D3_S1_L001_R1_002.fastq.gz
4 ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG004_NA24143_mother/NIST_Illumina_2x250bps/reads/D3_S1_L001_R1_003.fastq.gz
5 ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG004_NA24143_mother/NIST_Illumina_2x250bps/reads/D3_S1_L001_R1_004.fastq.gz
6 ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG004_NA24143_mother/NIST_Illumina_2x250bps/reads/D3_S1_L001_R1_005.fastq.gz
7 ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG004_NA24143_mother/NIST_Illumina_2x250bps/reads/D3_S1_L001_R1_006.fastq.gz
8 ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG004_NA24143_mother/NIST_Illumina_2x250bps/reads/D3_S1_L001_R1_007.fastq.gz
9 ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG004_NA24143_mother/NIST_Illumina_2x250bps/reads/D3_S1_L001_R1_008.fastq.gz
10 ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG004_NA24143_mother/NIST_Illumina_2x250bps/reads/D3_S1_L001_R1_009.fastq.gz
11 ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG004_NA24143_mother/NIST_Illumina_2x250bps/reads/D3_S1_L001_R1_010.fastq.gz
12 ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG004_NA24143_mother/NIST_Illumina_2x250bps/reads/D3_S1_L001_R1_011.fastq.gz
13 ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG004_NA24143_mother/NIST_Illumina_2x250bps/reads/D3_S1_L001_R1_012.fastq.gz
14 ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG004_NA24143_mother/NIST_Illumina_2x250bps/reads/D3_S1_L001_R1_013.fastq.gz
15 ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG004_NA24143_mother/NIST_Illumina_2x250bps/reads/D3_S1_L002_R1_001.fastq.gz
16 ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG004_NA24143_mother/NIST_Illumina_2x250bps/reads/D3_S1_L002_R1_002.fastq.gz
17 ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG004_NA24143_mother/NIST_Illumina_2x250bps/reads/D3_S1_L002_R1_003.fastq.gz
18 ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG004_NA24143_mother/NIST_Illumina_2x250bps/reads/D3_S1_L002_R1_004.fastq.gz
19 ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG004_NA24143_mother/NIST_Illumina_2x250bps/reads/D3_S1_L002_R1_005.fastq.gz
20 ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG004_NA24143_mother/NIST_Illumina_2x250bps/reads/D3_S1_L002_R1_006.fastq.gz
```

8 lines (7 sloc) | 2.11 KB

RawBlameHistory

```
BAM BAM_MD5 BAI BAI_MD5
ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/NIST_Illumina_2x250bps/novoalign_bams/HG002_GRCh38.2x250.bam 69f
ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/NIST_Illumina_2x250bps/novoalign_bams/HG002_hs37d5.2x250.bam 1f9
ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG003_NA24149_father/NIST_Illumina_2x250bps/novoalign_bams/HG003_GRCh38.2x250.bam 1b1
ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG003_NA24149_father/NIST_Illumina_2x250bps/novoalign_bams/HG003_hs37d5.2x250.bam 706
ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG004_NA24143_mother/NIST_Illumina_2x250bps/novoalign_bams/HG004_GRCh38.2x250.bam e08
```

[*help users download the specific dataset from the ftp site!](#)

giab_publications

This repository contains a list of publications pertinent to the Genome in a Bottle project

Publication describing data collected by GIAB for NA12878, the Ashkenazim trio, and the Chinese trio:

- Justin M Zook, David Catoe, Jennifer McDaniel, Lindsay Vang, Noah Spies, Arend Sidow, Ziming Weng, Yuling Liu, Chris Mason, Noah Alexander, Dhruva Chandramohan, Elizabeth Henaff, Feng Chen, Erich Jaeger, Ali Moshrefi, Khoa Pham, William Stedman, Tiffany Liang, Michael Saghbini, Zeljko Dzakula, Alex Hastie, Han Cao, Gintaras Deikus, Eric Schadt, Robert Sebra, Ali Bashir, Rebecca M Truty, Christopher C Chang, Natali Gulbahce, Keyan Zhao, Srinka Ghosh, Fiona Hyland, Yutao Fu, Mark Chaisson, Jonathan Trow, Chunlin Xiao, Stephen T Sherry, Alexander W Zaranek, Madeleine Ball, Jason Bobe, Preston Estep, George M Church, Patrick Marks, Sofia Kyriazopoulou-Panagiotopoulou, Grace Zheng, Michael Schnall-Levin, Heather S Ordonez, Patrice A Mudivarti, Kristina Giorda, Marc Salit, Genome in a Bottle Consortium, **Extensive sequencing of seven human genomes to characterize benchmark reference materials**, Scientific Data 3, Article number: 160025 (2016) doi:10.1038/sdata.2016.25. <http://www.nature.com/articles/sdata201625>
- Justin M Zook, David Catoe, Jennifer McDaniel, Lindsay Vang, Noah Spies, Arend Sidow, Ziming Weng, Yuling Liu, Chris Mason, Noah Alexander, Dhruva Chandramohan, Elizabeth Henaff, Feng Chen, Erich Jaeger, Ali Moshrefi, Khoa Pham, William Stedman, Tiffany Liang, Michael Saghbini, Zeljko Dzakula, Alex Hastie, Han Cao, Gintaras Deikus, Eric Schadt, Robert Sebra, Ali Bashir, Rebecca M Truty, Christopher C Chang, Natali Gulbahce, Keyan Zhao, Srinka Ghosh, Fiona Hyland, Yutao Fu, Mark Chaisson, Jonathan Trow, Chunlin Xiao, Stephen T Sherry, Alexander W Zaranek, Madeleine Ball, Jason Bobe, Preston Estep, George M Church, Patrick Marks, Sofia Kyriazopoulou-Panagiotopoulou, Grace Zheng, Michael Schnall-Levin, Heather S Ordonez, Patrice A Mudivarti, Kristina Giorda, Marc Salit, Genome in a Bottle Consortium, **Extensive sequencing of seven human genomes to characterize benchmark reference materials**, <http://biorxiv.org/content/early/2015/09/15/026468>.

Publication describing the methods used by NIST and GIAB to form v2.18 of the high-confidence SNP, indel, and homozygous reference calls for NA12878

- J.M. Zook, B. Chapman, J. Wang, D. Mittelman, O. Hofmann, W. Hide and M. Salit. **Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls**, Nature Biotechnology Published online Feb. 16, 2014. doi:10.1038/nbt.2835. PMID: [24531798](https://pubmed.ncbi.nlm.nih.gov/24531798/)

giab_FAQ

This repository contains FAQs (Frequently Asked Questions) for the Genome in a Bottle Consortium

1. If I have data or analyses from GIAB samples, how can I submit data to GIAB ftp site?

If you have data or analysis results to be submitted to GIAB ftp site, please contact Justin Zook (Justin.Zook@nist.gov) and Chunlin Xiao (xiao2@ncbi.nlm.nih.gov). An instruction will be sent to you regarding how to set up an uploading account.

2. Where can I find the latest sequence data, alignment data, analysis results by analysis group for the GIAB project?

All the data from the GIAB project are under <ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data>. The data are being organized by the individual trio (or sample) first, then by sequencing platforms. For a list of raw sequence files (eg fastq, fasta, hdf5, xsq, bnx format) or alignment files (eg bams, or xmap/cmap format) to download for your own analysis, you can go the repository of "giab_data_indexes" (https://github.com/genome-in-a-bottle/giab_data_indexes) to find the index file of interest with the particular sequencing platform.

The "analysis" folder under each trio (or sample) contains analysis results submitted by the analysis group with the sub-folder names consisting of: (1). the name of the analyzer or submitter, (2). sequencing technology or dataset(s), (3). type of variant to be analyzed, (4). analysis tool being used, and (5). date that serves as "version".

3. Where can I find high-confidence variant calls and bed files from the GIAB project?

The latest release and old versions are always under: <ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/>

4. Why do you have both high-confidence vcf and bed files?

The vcf files describe high-confidence variant calls, and the bed file describe high-confidence regions. If you wish to assess false positive rates, then you must use the bed file. Any variant calls you find in the high-confidence regions that are not in the high-confidence vcf are putative false positives.

5. What tool should I use to compare my SNP and small indel calls to the GIAB high-confidence calls?

The Global Alliance for Genomics and Health Benchmarking Team is currently developing standardized definitions for performance metrics and tools to calculate these metrics, including around complex variants that can have multiple correct representations in vcf. Three tools are currently being actively developed by the team:

vcfeval from Real Time Genomics (<http://realtimegenomics.com/products/rtg-core-non-commercial/>)

hap.py from Illumina (<https://github.com/sequencing/hap.py>)

vgraph from Kevin Jacobs (<https://github.com/bioinformed/vgraph>)

6. Can I access GIAB data from Amazon cloud?

Yes. All the GIAB data has been mirrored into Amazon S3, and the bucket name is `s3://giab`.

7. How can I get updates about new data, analyses, and conference calls of the GIAB Analysis Team?

Sign up for our GIAB Analysis Team google group at <https://groups.google.com/forum/#!forum/giab-analysis-team>. You can also sign up for general GIAB emails (e.g., workshop announcements) at <https://groups.google.com/forum/#!forum/genome-in-a-bottle>.

8. What should I know about how to use the GIAB calls appropriately?

(a) It is important to use the high-confidence bed file if you wish to assess false positive rates. (b) We highly recommend manually inspecting a subset of putative false positives and negatives to understand whether they actually are errors and what causes them. (c) The high-confidence calls and regions are biased towards easier variants and regions, so one should not assume performance metrics can be extended to types of calls and regions that are not assessed. This is especially critical for some clinical tests that are enriched for difficult variants or regions. (d) Stratification of performance into different types of variants and different

GIAB data on Amazon Cloud

Entire data from GIAB ftp site have been synchronized into AWS S3:

The s3 bucket name is: `s3://giab`

Acknowledgements

NCBI:

Steve Sherry
Jonathan Trow
Dima Beloslyudtsev

NIST:

Justin Zook
Marc Salit

And many others